Nucleic Acids Research, 2005, Vol. 33, Database issue ***D325–D328***
doi:10.1093/nar/gki008

VFDB: a reference database for bacterial virulence factors

Lihong Chen, Jian Yang, Jun Yu[1], Zhijian Yao[2], Lilian Sun, Yan Shen[2] and Qi Jin*

State Key Laboratory for Molecular Virology and Genetic Engineering, Beijing 100052, China, [1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK and [2]National Center of Human Genome Research, Beijing 100176, China

Received August 14, 2004; Revised and Accepted September 14, 2004

ABSTRACT

Bacterial pathogens continue to impose a major threat to public health worldwide in the 21st century. Intensified studies on bacterial pathogenesis have greatly expanded our knowledge about the mechanisms of the disease processes at the molecular level over the last decades. To facilitate future research, it becomes necessary to form a database collectively presenting the virulence factors (VFs) of various medical significant bacterial pathogens. The aim of virulence factor database (VFDB) (http://www.mgc.ac.cn/VFs/) is to provide such a source for scientists to rapidly access to current knowledge about VFs from various bacterial pathogens. VFDB is comprehensive and user-friendly. One can search VFDB by browsing each genus or by typing keywords. Furthermore, a BLAST search tool against all known VF-related genes is also available. VFDB provides a unified gateway to store, search, retrieve and update information about VFs from various bacterial pathogens.

INTRODUCTION

Despite advances in the prevention and treatment of infectious disease, pathogenic bacteria remain the pre-eminent threats to public health worldwide (1). Increasing antibiotic resistance strains and emerging and reemerging infectious agents cause alarming new concerns (2–5). Accordingly, the field of bacterial pathogenesis has rapidly expanded with a greater understanding of pathogenesis at the molecular level over the last decades. The term virulence refers to a quantitative measure of the pathogenicity or the likelihood of a pathogen causing infection (6). However, virulence factors (VFs) apply to the elements (i.e. gene products) that enable a microorganism to colonize a host niche where the organism proliferates and causes tissue damage or systemic inflammation. Conventional VFs include secreted proteins, such as protein toxins and enzymes, and cell-surface structures, such as capsular polysaccharides, lipopolysaccharides and outer membrane proteins, which directly contribute to the disease processes. Now, it becomes clear that many genes encoding virulence traits, such as secretion machineries, siderophores, catalases, regulators, etc. are indirectly involved in pathogenesis, which is equally important for bacteria to establish infection (7).

In recent years, rapid progress in bacterial genomic sequencing has led to the discovery and characterization of many new VFs from diverse species, resulting in numerous research papers and reviews each year. This, unfortunately, leads to scattered data at different places, including hard-copy publications as well as many web sites. Obtaining comprehensive information on each of the VFs becomes a formidable task. The virulence factor database (VFDB) reported in this communication is aimed at providing an in-depth coverage of the major VFs from various best-characterized bacterial pathogens, with emphasis on functional and structural biology, and essential immunology. This thorough and comprehensive VFDB will be maintained up to date, which would be helpful for researchers to elucidate pathogenic mechanisms in infectious diseases that require further immediate investigations for the development of novel approaches of the disease treatment and prevention.

The currently released VFDB contains cumulative information of VFs for 16 important bacteria pathogens, virulence-associated genes, protein structural features, functions, mechanisms and important literatures. Pathogenicity islands (PAIs), which are clusters of genes encoding virulence traits (8), are also included. Many graphic illustrations are made according to the original research papers or reviews for an easy grasp of the functions and structures of some VFs. The

*To whom correspondence should be addressed. Tel: +86 10 6787 7732; Fax: +86 10 6787 7736; Email: zdsys@sina.com

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

database is user-friendly; fully searchable by query text or by function categories of VFs, as well as by BLAST (9) comparison against all the VF-related genes.

SOURCE OF DATA AND DATABASE CONSTRUCTION

VFDB contains an extensive collection of VFs from 16 important bacterial pathogens. For each genus, those experimentally demonstrated VFs were collected first based on the original research papers appeared in PubMed to form the primary database. We then used Perl scripts to extract positions and sequences of these VFs from the GenBank (10) files. The Clusters of Orthologous Group (COG) (11) database was used to make classification for each gene record in the database if there lacks such information in their original annotation files. Additional information was obtained by searching NCBI website.

We built VFDB on a recent version of Linux operation system on a DELL PowerEdge™ 1600SC server, and MySQL was used to construct the background relational database to store the information of VFs and the gene sequences data. The Perl programming language and several modules, such as DBI, GD and CGI, were used to communicate with the database, dynamically generate the circular or linear maps and send all kinds of pages to the Apache web server for users. Some webpage effects were also made by client-side JavaScript.

HOW TO USE VFDB

VFDB can be freely accessed through a web browser at http://www.mgc.ac.cn/VFs/ for all researchers. In general, there are at least two ways to use the database: browse and search.

Browse in the database

In the main page of VFDB, a clickable list of currently available genus is organized at the left side, allowing users to browse pages for each of the genus. A typical first page for a genus is shown in Figure 1A. It begins with general information about the genus, such as nomenclature, classification, characteristics of the virulence mechanism and manifestation of the disease the bacterium causes. Following the general information is a summary of all available complete genomes and the corresponding publications with direct links to Entrez Genome browser or PubMed. For some genus, manually drawn graphic illustrations are used to illustrate their life cycles or/and most important pathogenic strategies. All the known major VFs are then listed in alphabetical order of function categories, with further information provided in subclass individual pages. The distribution of VFs in the genome is depicted in a circular map at the bottom of each page, which gives basic information about the genome in the center. Genes encoding VFs on leading and lagging strands are separately presented as clickable bars in the map, with colored code for different COG classifications. The existing PAI regions are also indicated in the map with links to individual PAI pages.

Individual VF is grouped into the function categories and is listed alphabetically, with detailed information about related genes, keywords, structure features, functions and mechanisms, as well as the original literatures or important reviews accessible through direct links to PubMed (for example see Figure 1B). In PAI pages, besides basic information about the PAI, such as location, size, G+C content, insertion locus, phenotype, remark and references, a linear map of the PAI region is also presented at the top of the page which shows the detailed organization of the PAI (Figure 1C). This linear map is color coded as those in the circular map in VFs page. Each arrow depicts a coding gene, which is a clickable link to its subclass page. The gene page of the subclass gives location, coding strand and product information for the gene, as well as accession number that is directly linked to the individual annotation file in GenBank and the classification code or COG number linked to the COG database in NCBI. Both the complete DNA and protein sequences of the gene were provided in text boxes which can be easily downloaded as FASTA format files by clicking the graphic buttons above the boxes (Figure 1D). For convenience, all the three types of pages (VFs, PAIs and genes) have graphic navigation buttons at the bottom of the pages to browse quickly (for example see Figure 1B).

Search the database

VFDB provides a powerful search engine for users to extract information from the database quickly through three different ways: (i) text search, (ii) BLAST search and (iii) VFs function category search.

The text search enables extracting general information, VFs, PAIs or gene name and product using any querying keywords. A single entry, that is instantly familiar to users of other internet search engines, is offered for alternative query words, separated by blanks, or complex phrases enclosed by double quotation marks. The querying results are displayed in an explicit table with each hit represented by a row containing corresponding genus, VFs, PAIs or gene names and a summary of characteristics, functions or products. Furthermore, each listed row in the output table provides links to the individual subclass pages, which highlight the querying keywords found in the page for users to swiftly locate them.

The standalone WWW-BLAST program is carefully integrated into VFDB to allow users performing sequence comparison against all sequences (nucleonic acid or amino acid) in the database using the BLAST algorithm. In addition to customized parameters for the program and cross links to the corresponding subject gene pages in the output, we implement graphical representation of the sequence search results that is helpful in parsing BLAST outputs. Furthermore, we also provide two-sequence alignment (12) and PSI/PHI BLAST search to help users analyzing their sequences in VFDB.

In VFDB, browsing by genus as mentioned above provides a hierarchical way to understand each known VFs in a genus, while searching VFs by function category provides another way to find out the occurrence of each kind of VFs in all genus. By using such a 'lateral' comparison, users gain a different view to better understand the function of VFs in bacteria. In the current version of VFDB, VFs are divided into four categories: offensive, defensive, nonspecific and regulatory which lead to 35 subclasses. Clicking each of the subclass names generates

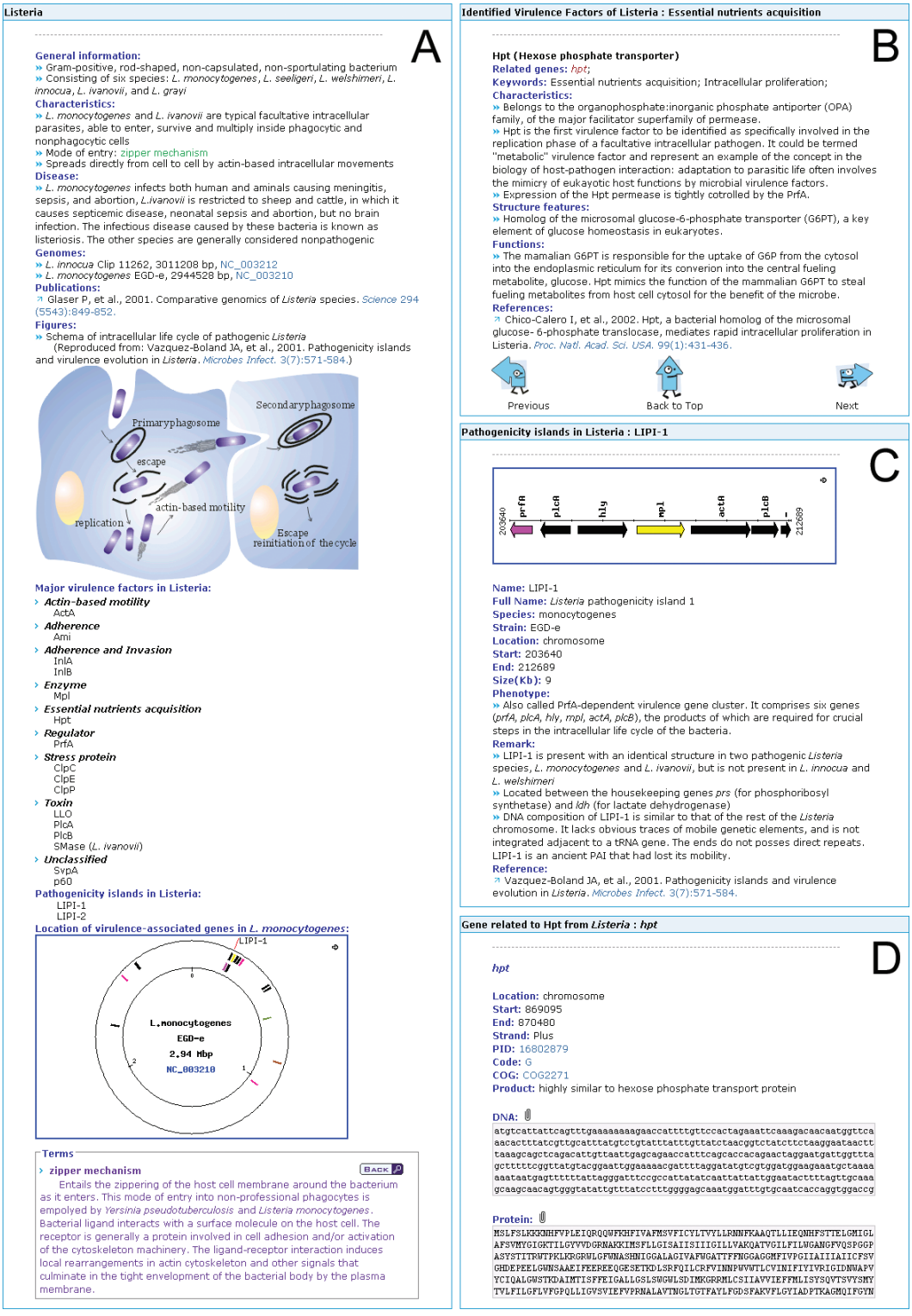

Figure 1. The main page contents of (**A**) the genus *Listeria*, (**B**) the *Listeria* hexose phosphate transporter, (**C**) the *Listeria monocytogenes* LIPI-1 pathogenicity island EGD-e and (**D**) the *Listeria hpt* gene page. The text in the box in the bottom of (A) explains the 'zipper mechanism' used by the organism during invasion.

a tabular result page similar to the output of 'text search' described above. Each genus and VFs names listed in the result page offers direct links to the corresponding genus or VFs pages in VFDB.

CURRENT STATUS AND FUTURE DIRECTIONS

VFDB is a specialized database aimed at making a comprehensive repository of known VFs complemented by data extraction and analysis tools to help further analysis of

Table 1. Detail statistics of VFs, PAI and related genes of each pathogen in VFDB (status of August 12)

Pathogens	Number of VFs	Number of PAIs	Related genes
Bacillus	2		7
Bordetella	11		72
Escherichia[a]	70	8	556
DAEC	1		5
EAEC	6		24
EHEC	20	1	66
EPEC	19	2	74
ETEC	3		11
MNEC	7		11
UPEC	14	5	365
Haemophilus	16		62
Helicobacter	12	1	59
Legionella	24		53
Listeria	16	2	18
Mycobacterium	34		62
Neisseria	12		50
Pseudomonas	20		171
Salmonella	17	5	174
Shigella	17	4	181
Staphylococcus	24		68
Streptococcus[a]	34		88
Streptococcus agalactiae	10		27
Streptococcus pneumoniae	10		26
Streptococcus pyogenes	14		35
Vibrio	6	2	82
Yersinia	10	3	123
Total	325	25	1826

[a]These genera have separate records according to different types or species as each possesses particular set of VFs.

VFs in bacteria. The current (August 12, 2004) statistics of VFDB is summarized in Table 1, and the most up-to-date status of the database is available at http://www.mgc.ac.cn/cgi-bin/VFs/status.cgi. We are committed to provide up-to-date VFs information by regular upgrading. Researchers are kindly invited and encouraged to deposit their new results of VFs at VFDB. Submission might either be performed through the 'Feedback' form accessible at the main page or by Email. Current VFDB contains information of VFs from 16 important bacteria, while more pathogens, such as *Chlamydia* and *Mycoplasm*, are under the way to be included into the database in the next release. As regulation of VFs expression is another important part of the virulence mechanisms, we plan to include more detailed information in this area in the future.

ACKNOWLEDGEMENTS

This project is supported by the National Basic Research Priorities Program and the High Technology Research and Development Program from the Ministry of Science and Technology of China.

REFERENCES

1. Waldvogel,F.A. (2004) Infectious diseases in the 21st century: old challenges and new opportunities. *Int. J. Infect. Dis.*, **8**, 5–12.
2. Byarugaba,D.K. (2004) Antimicrobial resistance in developing countries and responsible risk factors. *Int. J. Antimicrob. Agents*, **24**, 105–110.
3. Hogan,D. and Kolter,R. (2002) Why are bacteria refractory to antimicrobials? *Curr. Opin. Microbiol.*, **5**, 472–477.
4. Docampo,R. (2003) New and reemerging infectious diseases. *Emerg. Infect. Dis.*, **9**, 1030–1033.
5. Morens,D.M., Folkers,G.K. and Fauci,A.S. (2004) The challenge of emerging and re-emerging infectious diseases. *Nature*, **430**, 242–249.
6. Weiss,R.A. (2002) Virulence and pathogenesis. *Trends Microbiol.*, **10**, 314–317.
7. Brogden,K.A., Roth,J.A., Stanton,T.B., Bolin,C.A., Minion,F.C. and Wannemuehler,M.J. (2000) *Virulence Mechanisms of Bacterial Pathogens, 3rd edn.* ASM Press, Washington DC.
8. Hacker,J., Blum-Oehler,G., Muhldorfer,I. and Tschape,H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.*, **23**, 1089–1097.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
11. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
12. Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.